첨부 인공지능안전연구소 평가 실적

□ 2025년도 평가 모델 상세 정보 - 모델 29종

번호	평가 일자	모델명	평가 항목
1	2025.01	DeepSeek R1 0528	비물리적/물리적 위해, 무기확산, 민감정보유출, 사이버공격, 사회불안 및 정보검열
2	2025.06	Kanana SafeGuard Base 8B	윤리성 및 신뢰성 보장, 정보유출방지, 사이버공격, CBRN-E, 국가안보, 민감콘텐츠, 과도거부측정
3		Kanana SafeGuard Siren 8B	
4		Kanana PromptGuard 2.1B	
5		Granite Guardian 3.3 8B	
6		LlamaGuard2 8B	
7		LlamaGuard3 8B	
8		Llama PromptGuard 86M	
9		ShieldGemma 9B	
10		Granite Guardian 3.2 5B	
11	2025.08	Kakao Kanana 9.8B	살상무기, 사이버보안, 웹익스플로잇, 악성링크, 프롬프트주입, 응답위해성
12		Qwen2.5 7B	
13		Llama3.1 8B	
14		Mistral 7B	
15	2025.08	GLM 4.5 Air	비물리적/물리적 위해, 무기확산, 민감정보유출, 사이버공격, 사회불안 및 정보검열
16		Kimi K2	
17		Qwen3 235B 2507	
18		Llama4 Scout	
19		DeepSeek V3	
20	2025.10	GPT-5-mini	콘텐츠 안전 및 윤리성 보장, 정보 유출 방지 및 보안성 보장
21		GPT-4o-mini	
22		Grok-4-fast	
23		Gemini-2.5-flash	
24		Claude-3.5-haiku	
25	2025.11	LlamaGuard 4 12B	윤리성 및 신뢰성 보장, 정보유출방지, 사이버공격, CBRN-E, 국가안보, 민감콘텐츠, 과도거부측정
26		Llama PromptGuard 2 86M	
27		ShieldGemma2 4B	
28		Qwen3Guard Gen 8B	
29		Qwen3Guard Gen 0.6B	

□ 2026년도 평가 모델 상세 정보 - 모델 13종

번호	평가 일자	모델명	평가 항목
1	2026.01	GPT-5	에이전트 민감정보 유출 위험
2		Kimi K2 Thinking	
3		Mistral Small 24B	
4	2026.03	Claude Opus 4.6	에이전트 프롬프트 주입 공격 위험, 멀티 에이전트 환경에서의 민감정보 유출 / 연산 자원 남용 / 악성 행동 확산 위험
5		Claude Sonnet 4.6	
6		GPT-5.2	
7		Kimi K2.5	
8		DeepSeek V3.1	
9		Qwen3.5	
10	2026.04	Gauss 2 37B	살상무기, 사이버보안, 웹익스플로잇, 악성링크, 프롬프트주입, 응답위해성
11		Gemma 3 27B	
12		EXAONE 4.0.1 32B	
13		Qwen3 32B	

Att. Korea AISI Evaluation Results

□ Detailed Information on Models Evaluated in 2025 - 29 Models

No.	Date	Model	Evaluation Areas
1	2025.01	DeepSeek R1 0528	Biological/Chemical Harm, Weapon Proliferation, Sensitive Information Leakage, Cybersecurity Attacks, Social Harm & Information Integrity
2	2025.06	Kanana SafeGuard Base 8B	Ethical & Reliability Assurance, Information Leakage Prevention, Cybersecurity, CBRN-E, National Security, Harmful Content, Over-refusal Assessment
3		Kanana SafeGuard Siren 8B	
4		Kanana PromptGuard 2.1B	
5		Granite Guardian 3.3 8B	
6		LlamaGuard2 8B	
7		LlamaGuard3 8B	
8		Llama PromptGuard 86M	
9		ShieldGemma 9B	
10		Granite Guardian 3.2 5B	
11	2025.08	Kakao Kanana 9.8B	Lethal Weapons, Cybersecurity, Jailbreak/Exploit, Malicious Links, Prompt Injection, Response Harmfulness
12		Qwen2.5 7B	
13		Llama3.1 8B	
14		Mistral 7B	
15	2025.08	GLM 4.5 Air	Biological/Chemical Harm, Weapon Proliferation, Sensitive Information Leakage, Cybersecurity Attacks, Social Harm & Information Integrity
16		Kimi K2	
17		Qwen3 235B 2507	
18		Llama4 Scout	
19		DeepSeek V3	
20	2025.10	GPT-5-mini	Content Safety & Ethical Assurance, Information Leakage Prevention & Security Assurance
21		GPT-4o-mini	
22		Grok-4-fast	
23		Gemini-2.5-flash	
24		Claude-3.5-haiku	
25	2025.11	LlamaGuard 4 12B	Ethical & Reliability Assurance, Information Leakage Prevention, Cybersecurity, CBRN-E, National Security, Harmful Content, Over-refusal Assessment
26		Llama PromptGuard 2 86M	
27		ShieldGemma2 4B	
28		Qwen3Guard Gen 8B	
29		Qwen3Guard Gen 0.6B	

□ Detailed Information on Models Evaluated in 2026 - 13 Models

No.	Date	Model	Evaluation Areas
1	2026.01	GPT-5	Sensitive Information Leakage Risks in Agentic AI Systems
2		Kimi K2 Thinking	
3		Mistral Small 24B	
4	2026.03	Claude Opus 4.6	Prompt Injection Risks in Agentic AI Systems, Sensitive Information Leakage in Multi-Agent Environments, Compute Resource Abuse, Malicious Behavior Propagation
5		Claude Sonnet 4.6	
6		GPT-5.2	
7		Kimi K2.5	
8		DeepSeek V3.1	
9		Qwen3.5	
10	2026.04	Gauss 2 37B	Lethal Weapons, Cybersecurity, Jailbreak/Exploit, Malicious Links, Prompt Injection, Response Harmfulness
11		Gemma 3 27B	
12		EXAONE 4.0.1 32B	
13		Qwen3 32B	